Theory

Proportional-hazards models and intra-class correlation

Cox proportional-hazards models (or Cox regression) estimate the hazard (rate) of an event occurring in relation to predictor variables with time-to-event data (i.e., time taken till the event or censoring; Cox 1972). The hazard is a rate (or risk) of which an event occurring at time t. The hazard rate is defined in a Cox model as:

$$\lambda_i(t) = \lambda_0(t) \exp(b_1 x_1 + b_2 x_2 + ... + b_m x_m), \tag{1}$$

where $\lambda(t)$ is the hazard rate at time t for the ith subject (individual), $\lambda_0(t)$ is the baseline hazard rate, $\beta_1, \beta_2, ..., \beta_m$ are the regression coefficients, and $x_1, x_2, ..., x_m$ are the predictor variables. Note $\lambda_0(t)$ take the place of the intercept as $\exp(\ln(\lambda_0(t)) + b_1 x_1 + ...)$ or $\exp(b_0 + b_1 x_1 + ...)$ where $\ln(\lambda_0(t)) = b0$ ('ln' is a natural logarithm). Equation 1 can be rearranged to:

$$\ln\left(\frac{\lambda_i(t)}{\lambda_0(t)}\right) = b_1 x_1 + b_2 x_2 + ... + b_m x_m, \tag{2}$$

for the right hand side to take a linear form, which is more familiar for many readers although it does not have the intercept (i.e., b_0) and the residual term (i.e., ε). To fit such a model using, for example, R, one needs to provide the time-to-event data in the form of a `Surv` object (e.g., `Surv(time, event)` where `time` is time taken till an event or censoring and `event` is usually 0 or 1, indicating whether the event occurred or not). The Cox model can be fitted using the `coxph` in the `survival` package (Therneau et al. 2015).

Now let us assume that we have a single predictor variable, sex, x_{sex} for a time-to-event data set (e.g. a latency to solve a task), and we have a single random effect (or cluster) α (e.g. individuals or populations). The Cox proportional-hazards model can be extended to include a random effect (individual identify), which is often referred to as the 'frailty' term (Cox regressions with a single random effect is known as the frailty model):

$$\ln\left(\frac{\lambda_{ij}(t)}{\lambda_0(t)}\right) = b_{sex}x_{sex} + \alpha_i, \tag{3}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2),$$

where $\lambda_{ij}(t)$ is the hazard rate at time t for the ith individual for the jth occasion (observation). This model (frailty model) can be fitted using the `coxme` function in the R package `coxme` (Therneau 2015) as well as `coxph`.

Now we have defined the Cox model so let us define repeatably or intra-class correlations (ICC) in its simplest form when the trait of interest (the response variable) is a Gaussian variable (i.e. having normally distributed residuals):

$$\text{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}, \tag{4}$$

where σ_α^2 is the variance of the random effect (identity for a cluster: e.g. individuals so the between-cluster variance) and σ_ϵ^2 is the variance of the residuals (or within-cluster variance; Nakagawa & Schielzeth 2010). The ICC can be interpreted as the proportion of the total variance that is due to the between-cluster variance. The ICC can be calculated for (generalized) linear mixed-effect models (LMMS or GLMMs). For example, the R pacakge, `rptR` can be used to calculate the ICC for a GLMM, via `lmer` and `glmer` function in the `lme4` package.

Nakagawa & Schielzeth (2010) suggest that for non-Guassian data (e.g. binomial or Poisson), the within-cluster variance can be determined by what distributional assumptions GLMM makes (e.g. binomial or Poisson). For example, to obtain ICC for binary GLMMs on the latest (link/transformed) scale, σ_ε^2 can be assumed to be $\pi^2/3$ (they call σ_ε^2 as the distributional specific variance; for more details, see Nakagawa & Schielzeth 2010. Nakagawa et al. 2017). However, Cox models make no distributional assumptions about the hazard rate (i.e., non-parametric; Equation 1 & 2) although frailty models (Equation 3) has a random effect of a Gaussian distribution (this is why, this model is known as semi-parametric. Therefore, we could not calculate the ICC for Cox models.

Yet, in the statistical literature, a formula for parametric version of ICC (ICC_{np}) for the frailty model is known when we make the random effect is disbursed as a Gamma distribution on the exponential scale. If we denote the variance from a Gamma distribution as θ_α under Equation 3, ICC_{np} can be written as:

$$\text{ICC}_{\text{np}} = \frac{\theta_\alpha}{\theta_\alpha + 2}, \tag{5}$$

$$\exp(\alpha_i) \sim G\left(\frac{1}{\theta_\alpha}, \frac{1}{\theta_\alpha}\right),$$

where the first $1/\theta_\alpha$ and the second are $1/\theta_\alpha$, the shape and the rate parameter of the Gamma distribution, respectively (such parameterization results in the mean, $\mathrm{E}(\exp(\alpha_i)) = 1$ and variance, $\mathrm{Var}(\exp(\alpha_i)) = \theta_\alpha$).

The estimate, $\mathrm{ICC}_{\mathrm{np}}$ is Kendall's τ (the rank correlation or concordance for within-cluster observations) for the frailty model (Hougaard 2000). Unfortunately, there is no closed formula when assuming a Gaussian distribution for the random effect as in Equation 3. Yet, $\mathrm{ICC}_{\mathrm{np}}$ can be obtained numerically and we provide an R function based on the `tau` function from the R package, `parfm` (Munda et al. 2012). We note that and are unlikely to be the same but $\mathrm{ICC}_{\mathrm{np}}$ values, which we show in the simulation below.

An issue with the $\mathrm{ICC}_{\mathrm{np}}$ is that it is not a parametric version of ICC and more importantly, it is not clear whether this method can be extended where a Cox model has more than one random effect (at least, piratically speaking). Therefore, we need to use a trick to turn a time-to-event data for Cox models into a data set where we could fit a GLMM so that we can obtain parametric versions of ICCs.

Cox proportional-hazards models and generalized linear mixed models

In the statistical literature, it seems to be well known that the frailty model (Equation 3) can be fitted as a Poisson GLMM (known as, the piece-wise exponential model; e.g., Hirsh et al. 2016) or a binomial GLMM (the discrete-time model; Finkelstein 1986; for an accessible non-technical account, see Austin 2017). Here, we show the discrete-time model, more specifically, the binomial GLMM with the complementary log-log (cloglog) link can be used to fit a comparable model as Equation 3 by "exploding" the time-to-even data set by defining arbitrary discrete time intervals (Fig XXX shows an example of such a exploded data set compared to the original). If we assume we have three (arbitrary discrete) time intervals (*t1*, *t2* & *t3*) This binomial GLMM (without the intercept) is defined as:

$$\ln\left(\frac{-\ln(1-\lambda_{ijk}(t))}{-\ln(1-\lambda_{0k}(t))}\right) = b_{sex}x_{sex} + b_{t1}x_{t1} + b_{t2}x_{t2} + b_{t3}x_{t3} + \alpha_i, \tag{6}$$

where $\lambda_{ijk}(t)$ is the hazard rate at time t for the ith subject for the $j$the occasions for the kth time interval ($k = t1, t2, t3$), $\lambda_{0k}(t)$ is the baseline hazard rate for k is the time interval, x_{t1}, x_{t2}, x_{t3} are the indicator variables for the time intervals, and b_{t1}, b_{t2}, b_{t3} are the regression coefficients for the time intervals. Note that the cloglog link is $\ln(-\ln(1-p))$ where p is the probability of the event occurring so that the left hand side of Equation 6 consists of the cloglog-transformed hazard rate ($\lambda_{ijk}(t)$) and base hazard rate ($\lambda_{0k}(t)$).

Rather remarkably, b_{sex} and σ^2_α in Equation 6 are "the same" as those in Equation 3 although data structures used for two models are very different (i.e., time-to-event data vs. exploded data; Fig XXX). Using a simple simulation, we show the equivalence of b_{sex} and σ^2_α between the Cox (frailty) model (fitted with `coxph` and `coxme`) and the binomial GLMM (fitted with

`glmer` with `event` – binary data e.g., 0 or 1 – as the response. ; see below). Note that the number of intervals do not affect these estimates from the binomial GLMM.

Therefore, we can use variance components obtained from Cox models to estimate ICC under a binomial GLMM with the complementary log-log link where σ^2_ε (or the distributional-specific variance; Equation 4) is $\pi^2/6$ on the latent scale. This means we can define ICC for Equation 3 and 6 as (Nakagawa et al. 2017):

$$\text{ICC} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \pi^2/6}. \tag{7}$$

In the simulation below, we show the parametric version of ICC and the non-parametric version (ICC_{np}) are well correlated but not equivalent (analogous to the relationship between Pearson's r and Kendall's τ. We prefer the use of ICC as in Equation 7 over ICC_{np} because the parametric version is more comparable to other ICC estimates (on the latent/link scale) derived from GLMMs (e.g. Poisson and binomial data), which are now commonly used in ecology and evolution (Nakagawa et al. 2017).

Furthermore, the advantage of this approach is for us to add more than one random effect, which was eluded above. For example, imagine we have another cluster (a random effect), adding it to Equation 3 and then:

$$\ln\left(\frac{\lambda_{ijl}(t)}{\lambda_0(t)}\right) = b_{sex}x_{sex} + \alpha_i + \gamma_l, \tag{8}$$

$$\gamma_l \sim N(0, \sigma^2_\gamma),$$

where γ_l is the random effect for the lth level of the second cluster, which is normally distributed with the mean of zero and the variane of σ^2_γ. It is interesting to notice that the two random effects can be 'nested' or 'crossed' (Schielzeth and Nakagawa, 2013).

An example of the nested random effects are individual (α_i) and population (γ_l) where individuals are nested within populations. In this case, the ICC for individuals can be defined as:

$$\text{ICC}_{\text{ind1}} = \frac{\sigma^2_\alpha + \sigma^2_\gamma}{\sigma^2_\alpha + \sigma^2_\gamma + \pi^2/6}. \tag{9}$$

The reason both variance components are required for ICC for individuals is that some of similarities of a pair of individuals come from belonging to the same population. An example of the crossed random effects are individual (α_i) and year (γ_l) where individuals are not nested within years. This time, the ICC for individuals can be written as:

$$\text{ICC}_{\text{ind2}} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\gamma + \pi^2/6}. \tag{10}$$

If one wants to remove or adjusting the effect of year, and then, ICC for individuals simplifies to Equation 7. Speaking of adjusting, all the ICC formulas presented above represents 'adjusted' repeatablity (ICC; *sensu* Nakagawa and Schielzeth, 2010) because the effect of sex is accounted for in these models above. We can obtain 'unadjusted' repeatablity (ICC) by fitting the model without the fixed effect (sex), for example by changing Equation 3:

$$\ln\left(\frac{\lambda_{ij}(t)}{\lambda_0(t)}\right) = \alpha_i. \tag{11}$$

Importantly, σ^2_α and so ICC values should be larger obtained from this model (Equation 11) than those obtained from Equation 3 given the fixed effect explains non-zero variance. On the online page (www.github….xxxx), we show how to fit models and obtain ICC estimates, described above.

Simulation results

Note - I do not think we need a section for simulation???

Case studies